

# Query Expansion Using SNOMED-CT and Weighing Schemes

**Dawit Girmay, Afshin Deroie**

York University, Toronto Canada

## Abstract

*Despite all the advancements that have been made in the field of Information Retrieval, there are still so many challenges. These challenges are magnified when the information that is being retrieved is in a specialized domain such as healthcare. In order to tackle these challenges and encourage research in these domains, TREC (Text RETrival Conference) has instituted a Clinical Track in 2014. This paper is the result of participation in 2014 TREC Clinical Track. It entails the approach and the results that were obtained by utilizing Ontology to expand the original topics. Ontology was used in order to improve the quality of the terms present in the queries or topics, so that the queries are better structured, and they can better target documents of interest. The value that each term brings to the result was measured by way of weighing method algorithms in the retrieval system, BM25 and InL2c1. For this research, we have used SNOMED-CT along with UMLS Methathesaurus as our ontology in medical domain to expand the queries.*

## General Terms

Query Expansion, Ontology, relevance score

## Keywords

Weighing Methods, SNOMED-CT

## 1. Introduction and Motivation

The Clinical Track is instituted to encourage text retrievals in the area of diagnostics, treatment, and testing. The hope is that individuals in the healthcare field could leverage a retrieval model and make faster and sound decisions about the three areas of interest. As such, TREC has provided text documents from which participants need to use 30 topics and extract relevant documents. The documents provided are from Pub Med Central's database.

In the remainder of this paper, we discuss related works, give overview of the method of retrieval that we propose and instituted, discuss result of submission, and finally end with conclusion and future works.

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>NOV 2014</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2014 to 00-00-2014</b>	
4. TITLE AND SUBTITLE <b>Query Expansion Using SNOMED-CT and Weighing Schemes</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>York University ,Toronto Canada,</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES <b>presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).</b>					
14. ABSTRACT <b>Despite all the advancements that have been made in the field of Information Retrieval there are still so many challenges. These challenges are magnified when the information that is being retrieved is in a specialized domain such as healthcare. In order to tackle these challenges and encourage research in these domains, TREC (Text RETrival Conference) has instituted a Clinical Track in 2014. This paper is the result of participation in 2014 TREC Clinical Track. It entails the approach and the results that were obtained by utilizing Ontology to expand the original topics. Ontology was used in order to improve the quality of the terms present in the queries or topics, so that the queries are better structured, and they can better target documents of interest. The value that each term brings to the result was measured by way of weighing method algorithms in the retrieval system, BM25 and InL2c1. For this research, we have used SNOMED-CT along with UMLS Methathesaurus as our ontology in medical domain to expand the queries.</b>					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT  <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES  <b>9</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

## 2. Related Work

There are several methods to achieve retrieval of clinical documents of interest to users. However, most related researches available make use of query expansion, and therefore, that method was of interest to our team as well. Incidentally, we start the discussion regarding related work with publication that had to do with query expansion.

Jun Miao et al. [1], from York University describe both query expansion and enhancement of information retrieval model. The authors start with objective of retrieving patients from a collection that had/have a certain medical condition. BM25 was used to get baseline result. Then the performance of this model was compared with other models that are called York UMC2 which was semantically enriched methodology, York UMQ3 which combined Rocchio's feedback method with BM25's weighing method and York UMP4 which basically did the same thing as UMQ3, but added the idea of proximity of terms. The end result showed that the expansion of the queries that was augmented ended up outperforming the baseline. The authors finally recommend using the full capacity of the different Ontology that they used such as MeSH.

Martinez et al. [2] from University of Melbourne, Australia and University of the Basque country discuss their finding on query expansion using external sources headlined by Unified Medical Language System (UMLS), Wikipedia, and Dbpedia. The query was enriched with medical terms from these systems. At the same time, ICD descriptions which are publicly available were put in documents and run through Terrier to index them. Then they used 35 test queries from TREC 2011 to retrieve ICD codes. This was done to take advantage of the ICD codes present in the documents. Once the ranking was available from Terrier, they picked ICD codes deemed appropriate and retrieved documents. This method outperformed the method that uses external sources. In the end, authors conclude by suggesting that expanding query could also have negative impact on performance.

Koopman et al. [3] from Australian e-Health Research Centre also talk about query expansion by using Semantics. This was a very interesting approach in that they map both the query and the document using UMLS. Then, they map concepts from both the query and document to SNOMED-CT ontology. Once that was in place, they were able to match the queries to documents. By doing this, they were able to improve performance because with the SNOMED-CT ontology in place they were just not looking for key words, but concepts. This was a semantic approach. However, based on what they presented in the paper, it was not easy to tell how much advantage they take on the SNOMED-CT Ontology. It was simply stated that they are using it to map UMLS codes to SNOMED-CT without really delving too much into the relationships of concepts in the SNOMED-CT ontology.

Qi et al.[4] from NEC Labs America experiment with, expansion with UMLS concept. However, they suggested that the result was less than expected, and they went on with the submission only with the other methods by excluding UMLS expansion.

Henriksson et al.[5] in from Stockholm University and University of California, San Diego present an approach where they find new synonyms for preferred terms in SNOMED-CT using distributional similarity Methods. They use MIMIC-II database which is a large medical corpora in order to come up with the additional synonyms. Their objective was not information retrieval, but to identify synonyms of preferred terms. They were able to successfully do that.

### **3. Our Approach for Retrieval**

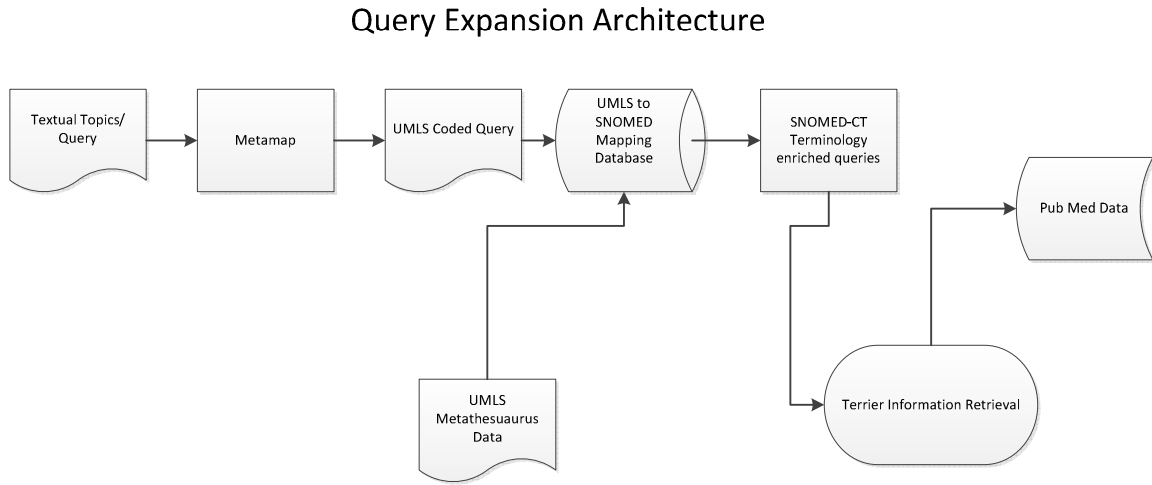
For the 2014 TREC clinical track, our research focuses on query expansion. The following are the details. The topics that have been provided are short clinical cases pertaining to three medical areas which are diagnosis, testing, and treatment. Then, document relevant to the case would need to be selected from PubMed Central (PMC) database file provided by TREC.

Keeping this in mind, the expansion intended in this research would use Metamap, UMLS Metathesaurus, and SNOMED-CT to find relevant documents pertaining to the query/topic.

The following are the proposed steps to achieve the retrieval set up.

- 1) The first step would be to map or enrich the textual query with Concept Unique Identifier(CUI) using Metamap tool for each of the thirty queries provided by TREC
- 2) From the Metamap results, terms associated with “Finding”, “Disease or Syndrome”, ”Sign Symptom” would be the ideal candidates for expansion
- 3) Then UMLS concepts generated from Metamap would be mapped to SNOMED-CT concepts using mapping file/database created using file provided by UMLS Metathesaurus
- 4) In the next step, SNOMED-CT concepts that are deemed appropriate, synonyms, based on relationships in the SNOMED-CT ontology, are retained in the query. These would be concepts that are plain English at this point.
- 5) Information retrieval system, Terrier from University of Glasgow School of Computing Science [6], would be used to do indexing and information retrieval using InL2C1 and BM25 weighing model

The following Diagram is the proposed architecture showing the interaction between the different systems involved in the retrieval process.



**Diagram 1 – This is a depiction of all the nodes involved for the retrieval process**

## Retrieval Process

In this section, a high level description of the retrieval process would be given. The objective of our retrieval process is to get clinical terms that would help us retrieve accurate and relevant documents. As such, we take a single query provided by TREC and get all possible terms that could be used to expand it. The query is then interjected with the new terms one at a time and the relevance score observed for movement upward or downward. Given the newly interjected term improves performance of the query, (upward trend), the new term is retained as part of the query. If not, we move to the next possible term in the queue and discard the term that did not help with performance.

The extent to which the newly interjected term contributes positively or negatively to performance is measured by looking at the relevance weight score of the first document assigned by the weighing scheme after the interjection of each expansion term.

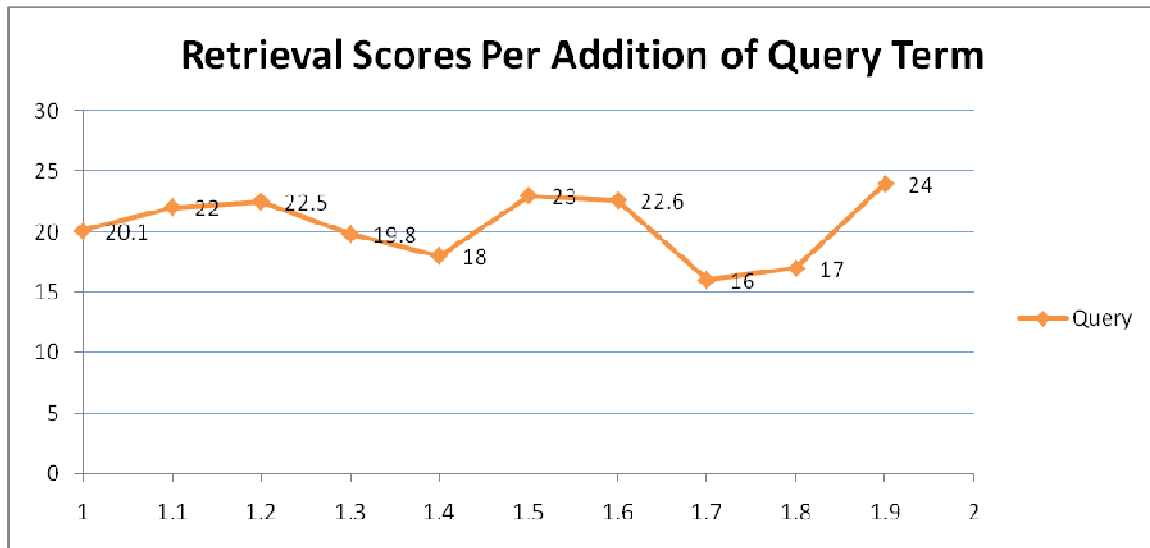
The following section is a graphical representation of how the expansion process works given query one from TREC 2014 and its sub queries.

Query	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Weight	20.1	22	22.5	19.8	18	23	22.6	16	17	24

**Table 1 – Relevance scores for added expansion terms**

The sub queries are generated when a possible expansion term has been identified that could be interjected into the original query. The assumption, in this case, is that there are

nine potential expansion terms as represented by the sub queries. The weight is relevance scores assigned to the first documents retrieved by any of the weighing schemes we used.



**Diagram 2 – Query Expansion values for a single query**

The x-axis represents the queries and y-axis, the retrieval score. The baseline is query #1.0 with score of 20.1. At the start, any term that would lower the score below that is discarded as we are looking to increase the score. Therefore, expansion terms that were in queries #1.1, #1.2, #1.5, #1.9 would be part of the final expanded query while the rest would not be. Note that query 1.6 would be discarded as its result would lower the score from query 1.5. To that end, it could be said that the cumulative score of all expansion terms at a point and time is what we focused on. Therefore, it's a moving baseline that starts from 20.1 and goes to 24 for query 1.9. Hence, given the above example, query 1.9 would be the final query as it would already be inclusive of all the prior expansion terms that improved performance.

## **Retrieval Types and Weighing Methods**

As briefly mentioned, Terrier Information Retrieval tool version 3.6 was used to retrieve the documents. Both the terrier default weighing method of InL2c1 and BM25 weighing methods were used. There were four types of retrievals that were done for our submission in 2014.

### **InL2c1 and BM25**

The summary queries that were provided by TREC were directly used to retrieve documents. For these two runs, there was no expansion of query involved. The summary part of the query was used to retrieve the documents, so these could be used as baseline.

## InL2c1EXP and BM25EXP

For these runs, the queries were expanded using the SNOMED-CT ontology. The results from the previous two runs of InL2c1 and BM25 were used as baseline for the expanded queries. Terms that were already in the summary queries that were also returned from SNOMED-CT were discarded as they would just be redundancies. Here is how the expansion process is initiated for the first query.

### Query #1

58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back

The first three terms that were identified in SNOMED-CT were:

*Hyperpiesia*

*Hypertensive*

*Systemic arterial hypertension*

Each one of the terms were then included with the original query as following.

### Query#1.0

*<Summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back</Summary>*

*<Expanded></Expanded>*

### Query #1.1

*<Summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back</Summary>*

*<Expanded> Hyperpiesia</Expanded>*

### Query #1.2

*<Summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back</Summary>*

*<Expanded> Hyperpiesia, hypertensive</Expanded>*

### Query #1.3

*<Summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back</Summary>*

*<Expanded> Hyperpiesia, Systemic arterial hypertension</Expanded>*

Note that Query #1.0 is always the same as the summary which is the original query provided by TREC. The next two sub queries shown above are the original query along with the potential expansion terms generated from SNOMED-CT.

Using the above example, we can assume that the addition of term “Hyperpiesia” onto the query gave a relevance weight score that was higher than its corresponding baseline for the first returned document. Hence, it was retained as part of the query that would make sub query 1.3. While on the other hand, query #1.2 was discarded by the time we got to query #1.3 as the term “hypertensive” was not present.

For our TREC submission, for topic one alone, there were 22 possible terms that were identified for the first query alone, so this process was repeated 22 times for all the terms that were extracted from SNOMED-CT. For each of the queries, the impact of the interjection of each of the terms was assessed by looking at the first relevance weight score. At the end of this recursive process, we were able to identify candidates from SNOMED-CT that would improve relevance. They were all then included as part of the query that was inclusive of all expansion terms. This query was used to do the retrieval for both InL2c1EXP and BM25EXP runs.

This process was repeated for all the thirty queries for all the possible SNOMED-CT terms that were yielded by the mapping database. The result of this process was submitted for TREC 2014 Clinical Track.

## 4. Results of Retrieval

Below table shows the result of the four submissions done as evaluated by TREC. All of our runs were manual. The following results pertain to InfAP, infNDCG, R-Prec, p@10 assessments.

Run	infAP	infNDCG	R-prec	P@10
BM25	0.0475	0.1763	0.1613	0.2667
BM25EXP	0.0477	0.1865	0.1625	0.2667
InL2c1	0.0475	0.1741	0.1582	0.2433
InL2c1EXP	0.0469	0.1865	0.1556	0.2800

**Table 2 Trec Evaluation Results**

Based, solely on InfAP, the performance of BM25Exp was the best of all the four. This is the query that had the original summary topics along with the expansion terms. The InL2c1 queries both the expanded and unexpanded versions performed lower than or same as the BM25 queries. This was a bit of a surprise to us, but not unexpected as we had anticipated InL2c1EXP would perform the best followed by BM25EXP. For InL2c1EXP, the interjection of expansion terms seems to have worsened InfAP.

Based on infNDCG, both the expanded queries BM25EXP and InL2c1EXP outperformed the unexpanded queries. This was in line with what we expected, and the expansion seems to have contributed positively in this respect. BM25EXP was also the superior performer when it comes to R-prec while both InL2c1 methods performed poorly. Finally InL2c1EXP had the best performance when it comes to P@10. This is very significant because our expansion mainly focused on the document that was ranked the highest having an improved score as a result of additional terms we introduced. It has also been our observation that the improvement of the score of the first file meant improvement in the weight score of the few files ranked right after the first one. Therefore, it seems like the terms introduced improved the mean precision of the first ten documents. Moreover, it is our belief that the first few files are what most users would be interested in when looking at returned results, so this was a significant finding.

Below is a table showing the breakdown for all the 11 manual runs for best and median for all the 2014 clinical track submissions.

Run	infAP	infNDCG	R-prec	P@10
Best	0.1308	0.3875	0.2586	0.5633
Median	0.0331	0.1615	0.1294	0.2367

**Table 3 TREC evaluation best and median results (Manual Runs)**

Below is a table showing the breakdown for all the 92 automatic runs for best and median for all the 2014 clinical track submissions.

Run	infAP	infNDCG	R-prec	P@10
Best	0.1805	0.5197	0.3496	0.7100
Median	0.0316	0.1514	0.1257	0.2333

**Table 4 TREC evaluation best and median results (Automatic Runs)**

Judging by this, all of the runs that we did performed better than the median in terms of all metrics infAP, infNDCG, R-Prec, p@10 for both manual as well as automatic runs.

## 5. Conclusions and Future Work

The clinical track of TREC, aims to help healthcare professionals retrieve documents that are related to diagnostics, treatment, and testing. To this end, we have constructed a simple but effective model that lets users do exactly that by using SNOMED-CT, UMLS, Metamap and lets users add query expansion terms. Once, terms were added, impact of terms on performance was monitored by looking at trend of weighing schemes on the first ranked document returned. For the purpose of our participation in 2014 TREC, we have submitted four runs that were able to perform reasonably well.

As a recommendation to expand this research, we propose looking into the different types of topics diagnostics, treatment, and testing and seeing if the retrieval would work better

for any particular genre. Secondly, although we noticed an improved score for the highest ranked document meant an improved score for the documents returned after the first one, we only looked at that there is an improvement on the highest ranked document's score as the main criteria. Perhaps, there could be a dynamic way where we can look at improvement of more documents rather than just the first one. This, in turn, could lead to overall improvement in terms of metrics used to measure retrieval precision. Last but not least, we believe using other well known or custom weighing methods could yield different results for the metrics. Therefore based, on the results we attained in our participation for 2014, we believe these ideas mentioned in the preceding lines deserve further look.

## References:

- [1] Jun Miao, Zheng Ye, Jimmy Huang, "York University at TREC 2012: Medical Records Track", TREC 2012
- [2] David Martinez, Arantxa Otegi, Eneko Agirre, "NICTA and UBC at the TREC 2012 Medical Track", TREC 2012
- [3] Bevan Koopman, Peter Bruza, Laurianne Sitbon, Michael Lawley "AEHRC & QUT at TREC 2011 Medical Track: a concept-based information retrieval approach", TREC 2011
- [4] Yanjun Qi and Pierre-Fran,cois Laquerre "Retrieving Medical Records with "sennamed":NEC Labs America at TREC 2012 Medical Records Track", TREC 2012
- [5] Aron Henriksson, Mike Conway, Martin Duneld, Wendy W. Chapman "Identifying Synonymy between SNOMED Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records", AMIA Annual Symposium Proceedings Archive, 2013
- [6] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). 10th August, 2006. Seattle, Washington, USA